

SUBDIRECCIÓN ACADÉMICA

DEPARTAMENTO DE SISTEMAS Y COMPUTACIÓN

ENERO - JUNIO 2020

INGENIERÍA INFORMÁTICA

MATERIA

DATOS MASIVOS

CATEDRÁTICO:

JOSE CHRISTIAN ROMERO HERNANDEZ

UNIDAD 2

# TAREA 2

ALUMNO

López Valencia Luis Daniel #15212159

### 

### **VectorAssembler** VectorAssembler es un transformador que combina una lista dada de columnas en una sola columna vectorial. Es útil para combinar características sin procesar y características generadas por diferentes transformadores de características en un solo vector de características, con el fin de entrenar modelos ML como la regresión logística y los árboles de decisión. VectorAssembler acepta los siguientes tipos de columnas de entrada: todos los tipos numéricos, tipo booleano y tipo vectorial. En cada fila, los valores de las columnas de entrada se concatenarán en un vector en el orden especificado. Ejemplo: Supongamos que tenemos un DataFrame con las columnas id, hour, mobile, userFeatures y clicked: userFeatures es una columna vectorial que contiene tres funciones de usuario. Queremos combinar las características de hora, móvil y usuario en un solo vector de características llamado características y usarlo para predecir clics o no. Si configuramos las columnas de entrada de VectorAssembler en hora, móvil y las características de usuario y la columna de salida en características, después de la transformación deberíamos obtener el siguiente DataFrame: **Distancia media cuadrática mínima (Root-mean-square deviation)** La raíz del error cuadrático medio (RECM) o raíz de la desviación cuadrática media (RDCM) es una medida de uso frecuente de las diferencias entre los valores (valores de muestra o de población) predichos por un modelo o un estimador y los valores observados. La RECM representa la raíz cuadrada del segundo momento de la muestra de las diferencias entre los valores previstos y los valores observados o la media cuadrática de estas diferencias. Estas desviaciones se denominan residuos cuando los cálculos se realizan sobre la muestra de datos que se utilizó para la estimación y se denominan errores (o errores de predicción) cuando se calculan fuera de la muestra. Formula